Techno International New Town

(Formerly Known As Techno India College Of Technology)

Department Of Computer Science & Technology

Project Tile:

Data Analytics On Diabetics Prediction

Team: Coding Monkeys

Members:

- *Rajarshi Baral*
- *Arpita Saha*
- *Sangeeta Barua*
- *Aoyan Mondal*

Under The Guidance Of
Prof. Mr. Swarup Chakraborty

--

INTRODUCTION

Diabetes is one of the most common chronic diseases affecting millions of people worldwide. It occurs when the body cannot effectively control blood sugar levels, which can lead to serious health complications if not detected and managed early. Early detection of diabetes can help individuals take preventive measures, receive timely treatment, and improve their quality of life.

With the rapid development of technology, machine learning (ML) has become a powerful tool in the healthcare sector. By analyzing historical medical data, ML algorithms can learn patterns and predict future outcomes with high accuracy. In the context of diabetes, these algorithms can help identify individuals at risk based on various characteristics such as age, BMI, blood pressure, insulin levels, and more . This project focuses on building a machine learning model that can accurately predict the likelihood of an individual developing diabetes.

PROBLEM DEFINATION

Feature	Description
Pregnancies	Number of times woman has been pregnant
Glucose	Plasma glucose concentration
Blood Pressure	Diastolic blood pressure(mm Hg)
Skin Thickness	Triceps skinfold thickness(mm)
Insulin	2-Hour serum insulin(mu U/ml)
BMI	Body Mass Index(weight in kg/(height in m)^2)
Diabetes Pedigree Function	A function that scores likelihood of diabetes based on family history
Outcomes	Class label: 0= no diabetes, 1 =diabetes

Objective - Gestational diabetes is a condition in which women without previously diagnosed diabetes exhibit high blood glucose levels during pregnancy. Early diagnosis is essential to manage health risks for both mother and child. Machine learning can help predict the likelihood of diabetes based on measurable health parameters.

- **Pre-existing Type 1 Diabetes:** Autoimmune condition where the body does not produce insulin. It is present before pregnancy.
- **Pre-existing Type 2 Diabetes:** A condition where the body becomes resistant to insulin or doesn't produce enough insulin. It is usually related to lifestyle and may be present before pregnancy.

PROBLEM DEFINATION

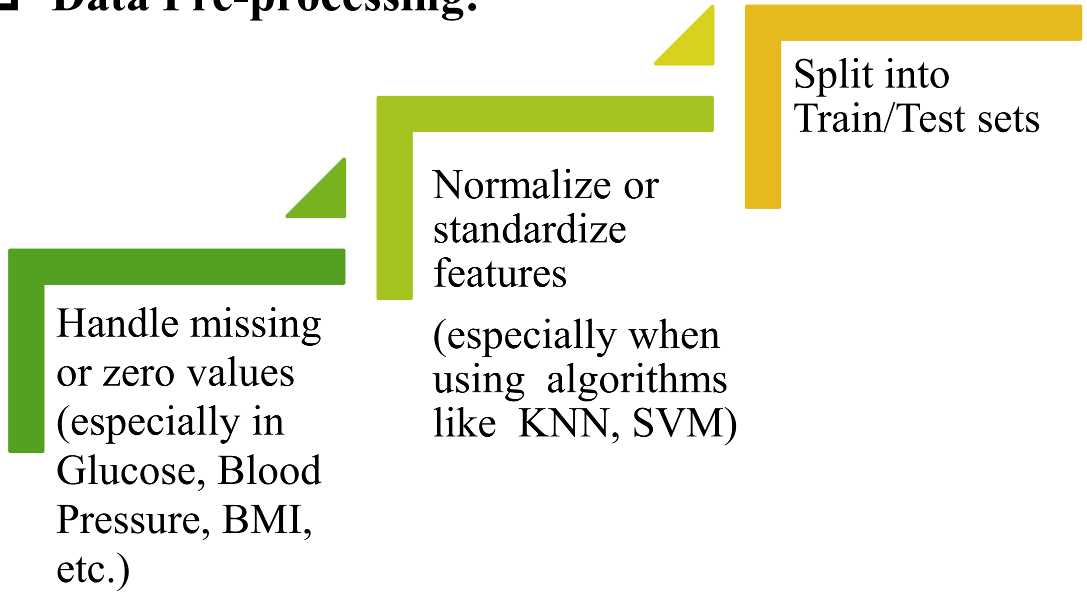
- How is the difference from others models/problems?

Aspect	This model	Typical Differences
Domain	Medical/Clinical	Other ML problems could be in finance, marketing , etc..
Data Type	Mostly Numeric	Some other models might include more categorical and text data
Feature Importance	All features are medically relevant	In generic ML tasks, feature relevant may not be so direct
Outcome Impact	High-stakes(medical)	Business models may be less critical in terms of error consequences.

PROBLEM DEFINATION

How to Solve It

❑ **Data Pre-processing:**



- Handle missing or zero values (especially in Glucose, Blood Pressure, BMI, etc.)
- Normalize or standardize features

  (especially when using algorithms like KNN, SVM)
- Split into Train/Test sets

❑ **Exploratory Data Analysis (EDA):**

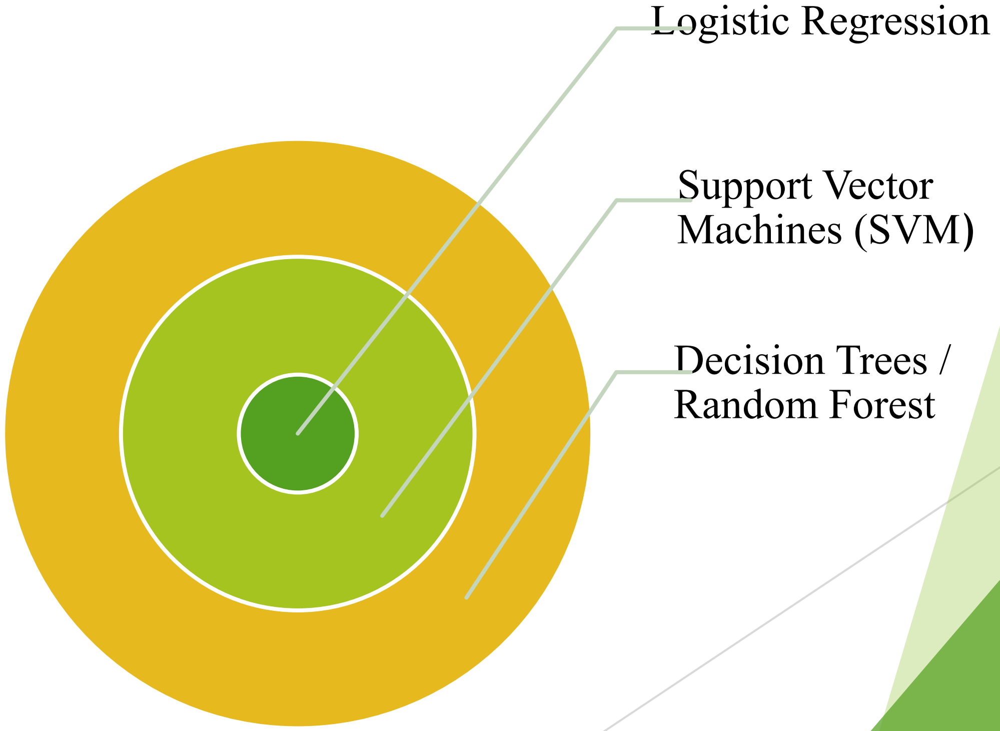
- **Visualize correlations**
- **Analyze feature distributions**
- **Understand class imbalance**

PROBLEM DEFINATION

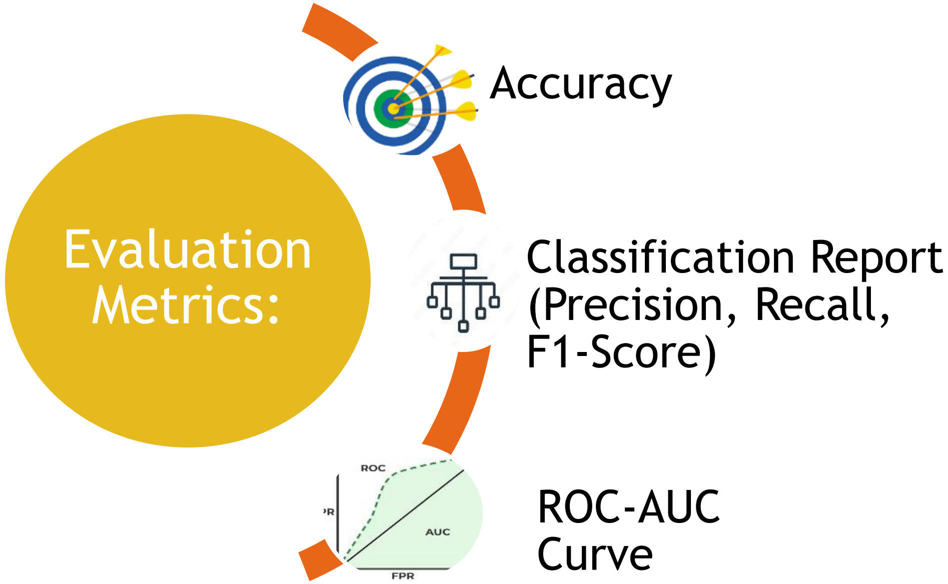
❑ Feature Selection :

Use techniques like Recursive Feature Elimination, PCA, or correlation analysis

❑ Model Selection:



PROBLEM DEFINATION



Evaluation Metrics:

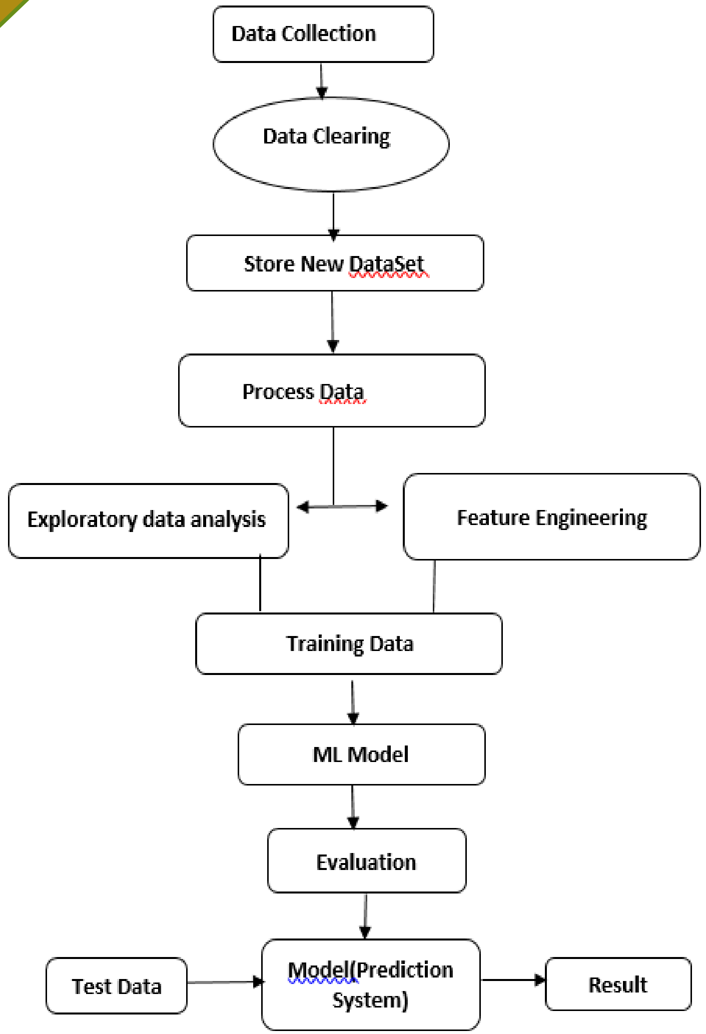
- Accuracy
- Classification Report (Precision, Recall, F1-Score)
- ROC-AUC Curve

□ Expected Output:

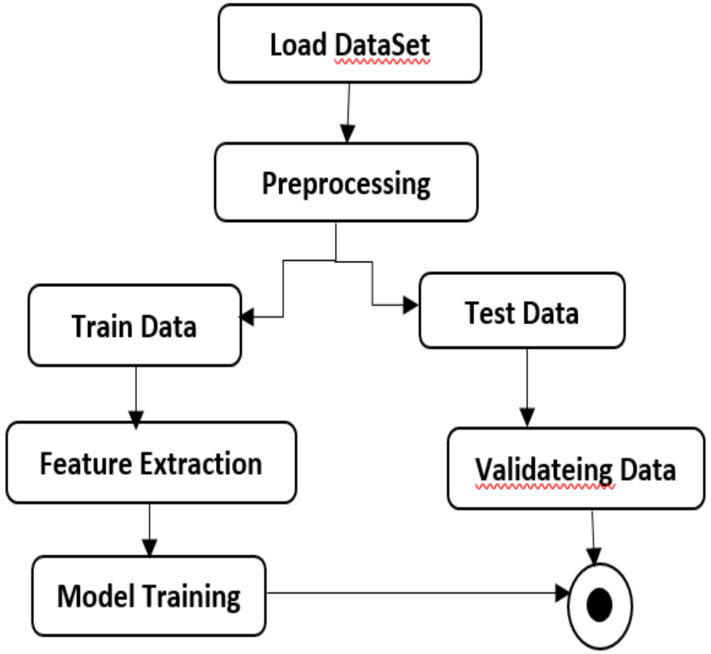
- A model that predicts the likelihood of gestational diabetes.
- A report on the most significant predictors .
- Visualization dashboards showing patterns, trends, and at-risk groups.

POGRESS

Flow chart

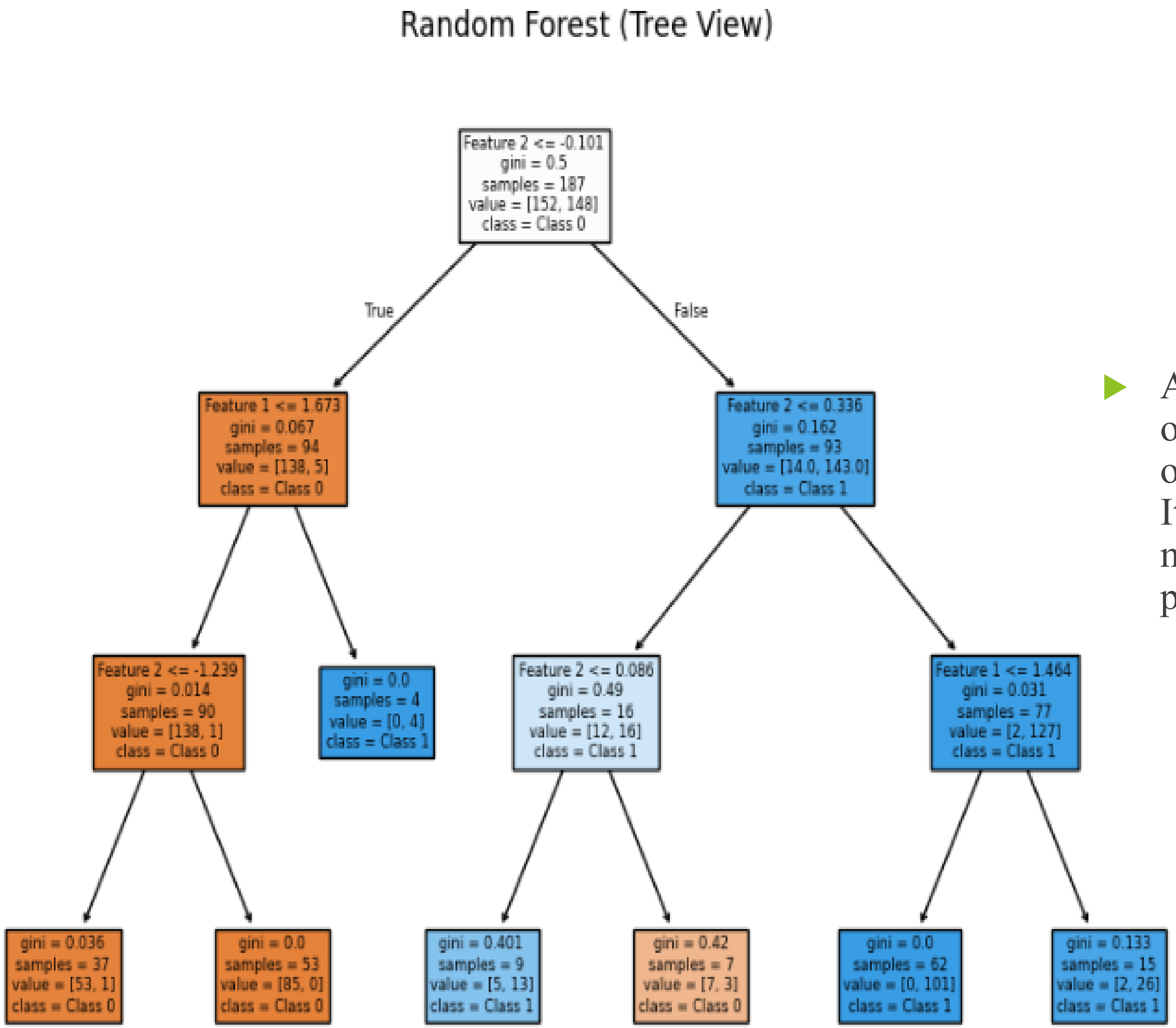


DIAGRAMS



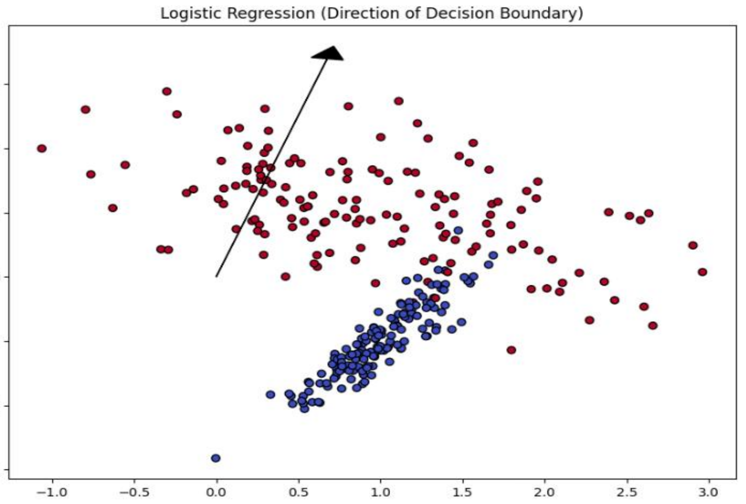
State diagram

PROGRESS



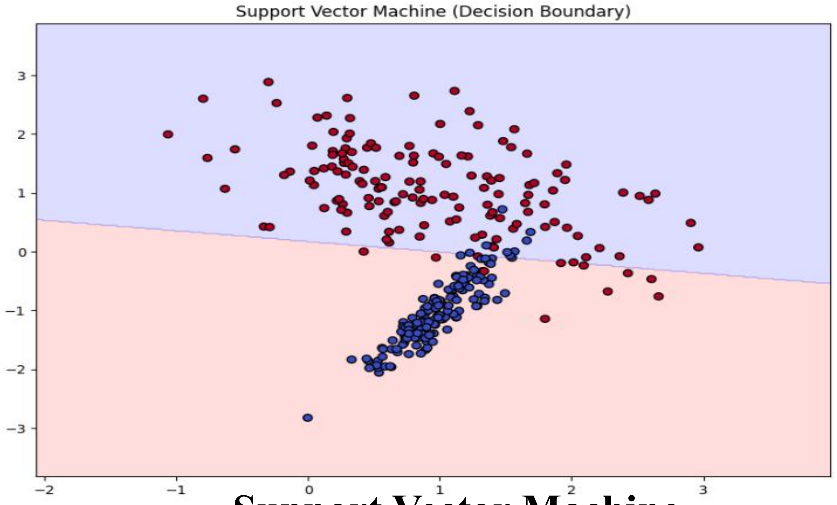
- A Random Forest is a collection of many Decision Trees trained on different parts of the dataset. It combines their results for a more accurate and stable prediction.

PROGRESS



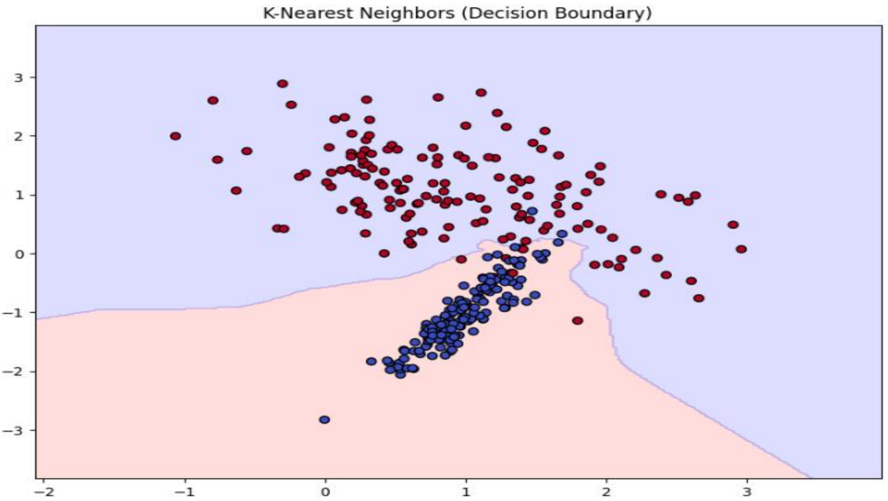
Logistic Regression

- ✓ Binary Prediction
- ✓ Understand which features impact the result
- ✓ Quick train and test



Support Vector Machine

- ✓ Handles Imbalance
- ✓ Robust to overfitting
- ✓ Versatile with kernels

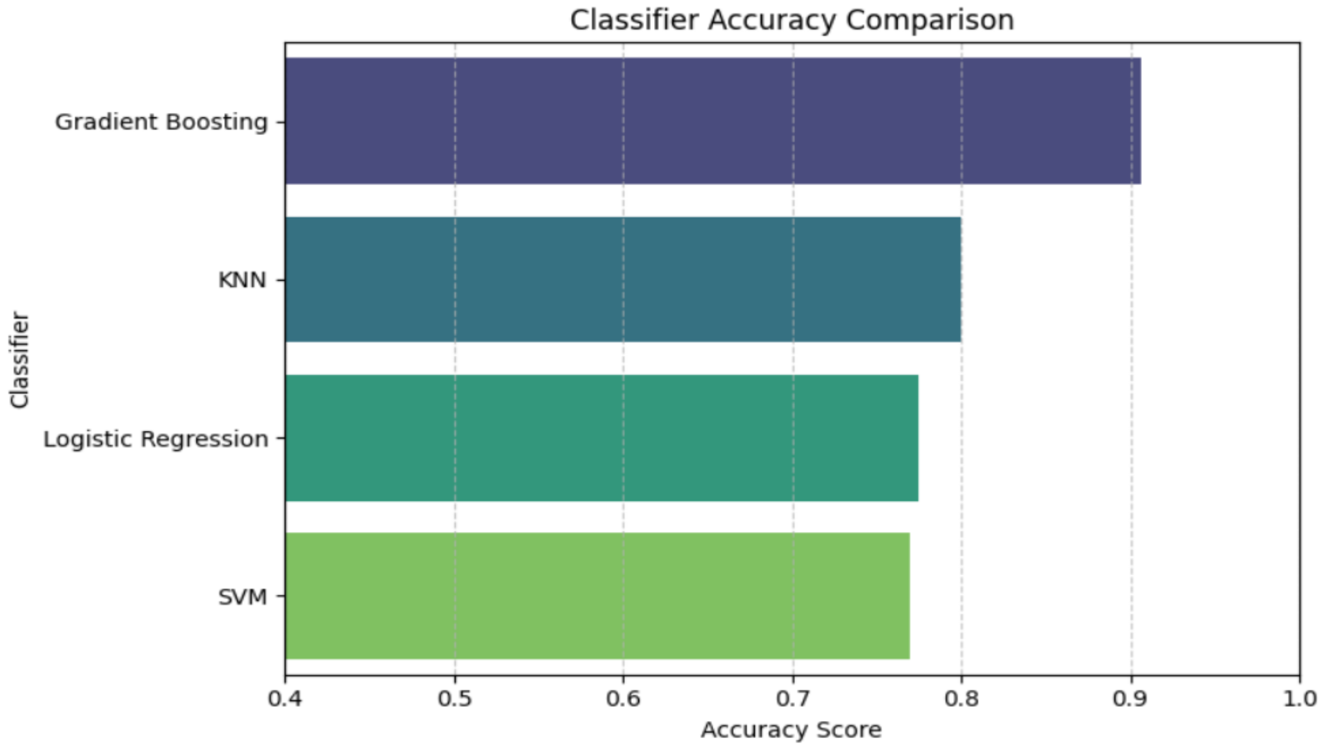


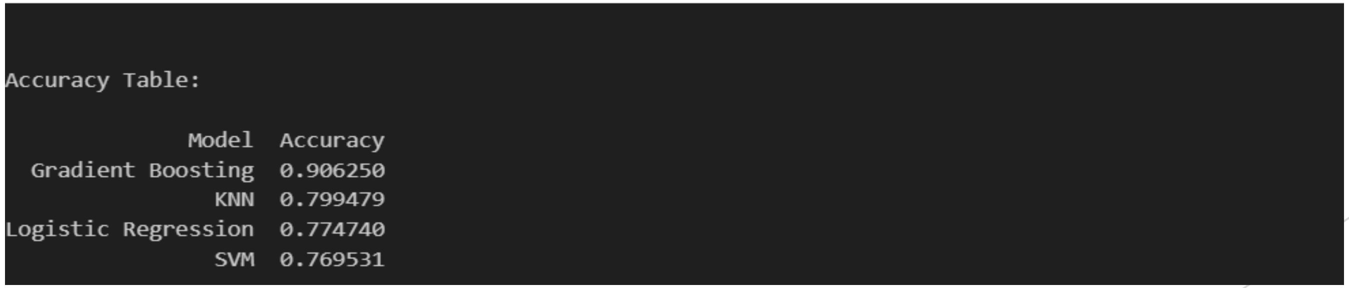
K-Nearest Neighbors

- ✓ Effective with normalized numeric features
- ✓ Easy to understand
- ✓ Sensitive to local patterns

PROGESS

Classifier Comparison

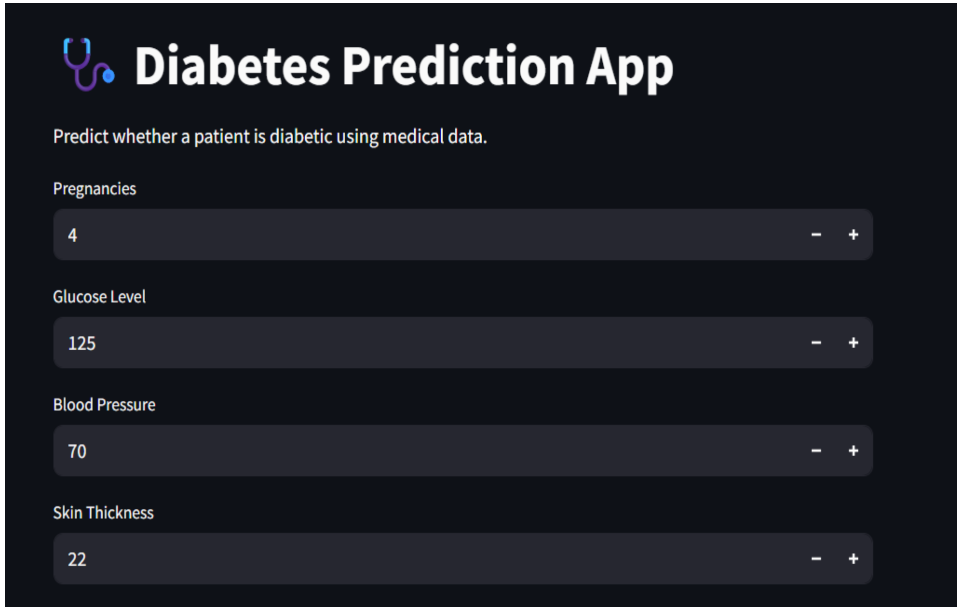




```
Accuracy Table:

                Model  Accuracy
   Gradient Boosting  0.906250
                  KNN  0.799479
Logistic Regression  0.774740
                  SVM  0.769531
```

RESULTS



🩺 Diabetes Prediction App

Predict whether a patient is diabetic using medical data.

Pregnancies

4

Glucose Level

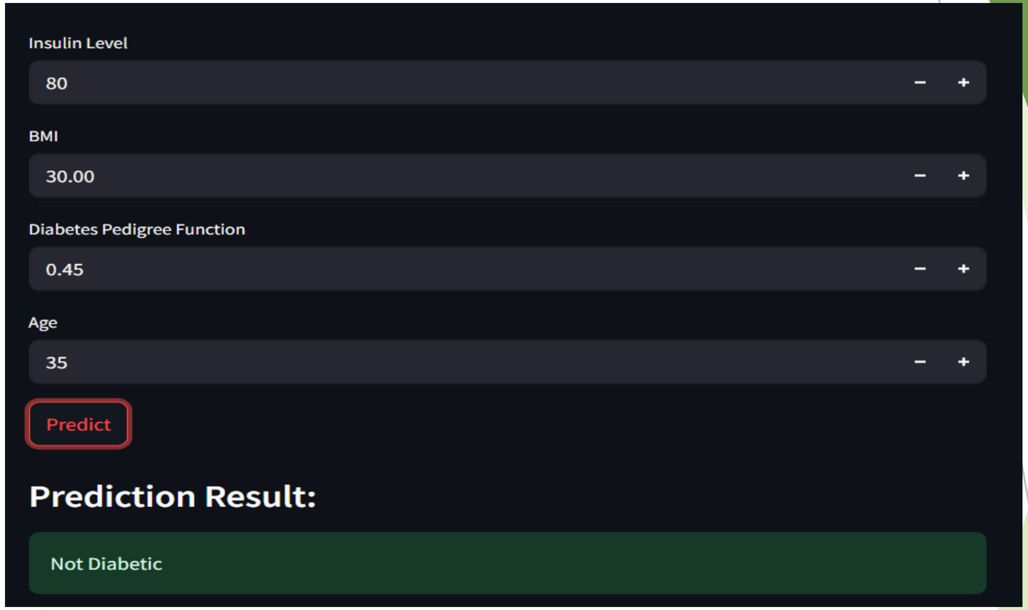
125

Blood Pressure

70

Skin Thickness

22



Insulin Level

80

BMI

30.00

Diabetes Pedigree Function

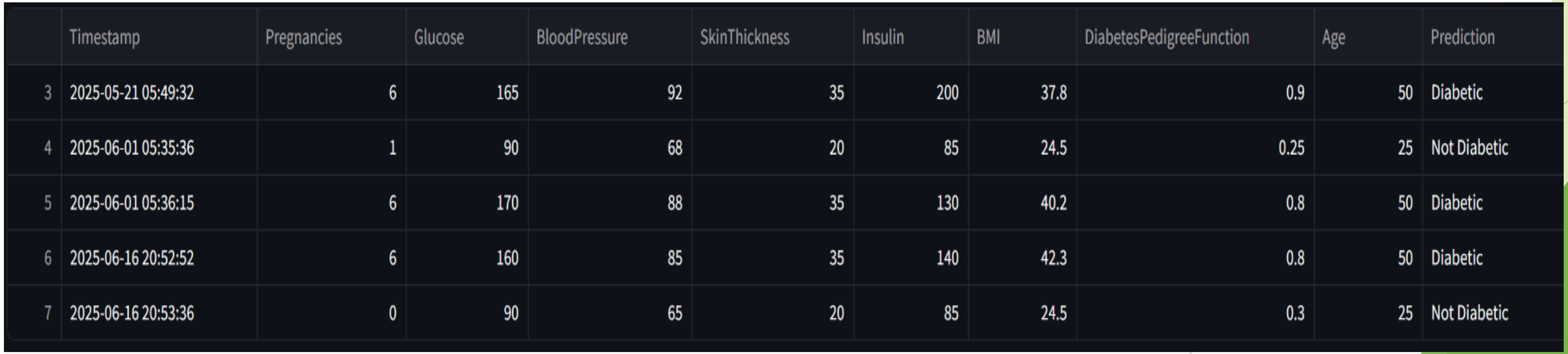
0.45

Age

35

Prediction Result:

Not Diabetic



	Timestamp	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Prediction
3	2025-05-21 05:49:32	6	165	92	35	200	37.8	0.9	50	Diabetic
4	2025-06-01 05:35:36	1	90	68	20	85	24.5	0.25	25	Not Diabetic
5	2025-06-01 05:36:15	6	170	88	35	130	40.2	0.8	50	Diabetic
6	2025-06-16 20:52:52	6	160	85	35	140	42.3	0.8	50	Diabetic
7	2025-06-16 20:53:36	0	90	65	20	85	24.5	0.3	25	Not Diabetic

FUTURE SCOPE

Advanced Predictive Analytics

Personalized Healthcare

Integration with Health Information Systems

Genomics and Biomarker Analytics

AI-Powered Decision Support Systems

CONCLUTION

Our Data Analytics on Diabetes Prediction project demonstrates the effectiveness of machine learning in identifying diabetes risk based on key health indicators. By preprocessing data, selecting relevant features, and applying predictive models, we achieve insightful results that can assist in early diagnosis and intervention. The correlation analysis helps understand relationships between variables, enhancing model interpretability. With further improvements in data quality and model optimization, this approach can significantly contribute to healthcare decision-making, promoting proactive diabetes management and awareness.

BIBLIOGRAPHY:

- **[1]** Sarwar, Muhammad Azeem, et al. "Prediction of diabetes using machine learning algorithms in healthcare." 2018 24th international conference on automation and computing (ICAC). IEEE, 2018.
- **[2]** Mujumdar, Aishwarya, and Vb Vaidehi. "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299.
- **[3]** Kumar, P. S., & Pranavi, S. (2017, December). Performance analysis of machine learning algorithms on diabetes dataset using big data analytics. In 2017 international conference on infocom technologies and unmanned systems (trends and future directions)(ICTUS) (pp. 508-513). IEEE.
- **[4]** Hassan, Md Mehedi, et al. "Early predictive analytics in healthcare for diabetes prediction using machine learning approach." 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 2021.
- **[5]** Kalyankar, Gauri D., Shivananda R. Poojara, and Nagaraj V. Dharwadkar. "Predictive analysis of diabetic patient data using machine learning and Hadoop." 2017 international conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC). IEEE, 2017.
- **[6]** Nibareke, Thérence, and Jalal Laassiri. "Using Big Data-machine learning models for diabetes prediction and flight delays analytics." Journal of Big Data 7.1 (2020): 78.

THANK YOU